The role of the production system in making predictions during comprehension

The role of language production in making predictions during comprehension

Jarosław R. Lelonkiewicz, Hugh Rabagliati, & Martin J. Pickering

Scuola Internazionale Superiore di Studi Avanzati, Trieste, Italy

University of Edinburgh, Edinburgh, UK

Author Note

This research was supported by a School Research Support Grant from the School of Philosophy, Psychology and Language Sciences, University of Edinburgh to J. R. L., and by a grant from the ESRC (ES/L01064X/1) to H. R.

Correspondence concerning this article should be addressed to Jarosław R. Lelonkiewicz, Cognitive Neuroscience Area, Scuola Internazionale Superiore di Studi Avanzati, via Bonomea 265, 34136 Trieste, Italy.

e-mail: jlelonki@gmail.com

Abstract

Language comprehension depends heavily upon prediction, but how predictions are generated remains poorly understood. Several recent theories propose that these predictions are in fact generated by the language production system. Here, we directly test this claim. Participants read sentence contexts that either were or were not highly predictive of a final word, and we measured how quickly participants recognized that final word (Experiment 1), named that final word (Experiment 2), or used that word to name a picture (Experiment 3). We manipulated engagement of the production system by asking participants to read the sentence contexts either aloud or silently. Across the experiments, participants responded more quickly following highly predictive contexts. Importantly, the effect of contextual predictability was greater when participants had read the sentence contexts aloud rather than silently, a finding that was significant in Experiment 3, marginally significant in Experiment 2, and again significant in combined analyses of Experiments 1-3. These results indicate that language production (as used in reading aloud) can be used to facilitate prediction. We consider whether prediction benefits from production only in particular contexts, and discuss the theoretical implications of our evidence.

keywords: prediction, language production, comprehension, simulation

The role of language production in making predictions during comprehension

1 Introduction

When reading or listening, people can predict the next word that they will see or hear (e.g., Altmann & Kamide, 1999; Federmeier & Kutas, 1999; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005). But what mechanisms do they use to do so? Recently, a number of researchers have proposed that comprehenders predict by involving aspects of the system that is otherwise used to produce utterances (prediction-by-production: Dell & Chang, 2014; Federmeier, 2007; Pickering & Garrod, 2013; cf. Hickok, 2012). More specifically, comprehenders may covertly imitate the utterance that they are currently hearing, and use this as the basis for determining what they themselves would say next if they were speaking.

However, the current evidence that production is used for prediction is largely indirect (see Pickering & Gambi, 2018). For instance, brain regions that are implicated in the production of tongue-articulated sounds are also active when comprehenders expect to hear such sounds (D'Ausilio, Jarmolowska, Busan, Bufalari, & Craighero, 2011), and tongue movements performed while listening to sentences appear to be affected by expectations about upcoming words (Drake & Corley, 2015). More recently, Rommers, Dell, and Benjamin (2020) had participants read sentence-final predictable or unpredictable words aloud or silently, and found that the production effect (the finding that words read aloud are remembered better than words read silently) was smaller for the predictable words. This implies that participants used the production system to predict such words and therefore reading aloud added less benefit to memory. Moreover, participants found it difficult to

remember whether a predictable word had been read aloud or silently, presumably because they engaged the production system in both cases.

All of these findings suggest a close link between production and prediction, but do not demonstrate that production is causally involved in prediction. There is also correlational evidence that people with better production skills are better at predicting language (Federmeier, Kutas, & Schul, 2010; Mani & Huettig, 2012), and some indirect evidence that prediction could be stronger in contexts where the production system is overall highly activated (Hintz & Meyer, 2015; Hintz, Meyer, & Huettig, 2016), but again, these studies do not unambiguously demonstrate a causal role of production in how individuals predict.

A causal test of prediction-by-production would show that intervening on the production system changes how participants generate predictions, and one recent ERP study provides this type of test. Martin, Branzi, and Bar (2018) had Spanish-speaking participants read highly predictive sentence contexts followed by either expected or unexpected words that differed in grammatical gender (e.g., *El rey llevaba en la cabeza una corona/un sombrero;* "The king wore on his head a crown/a hat"). It is known that an article or adjective whose gender is consistent with an unexpected but not an expected upcoming noun leads to an enhanced N400 effect, suggesting that the comprehenders predict such information (Van Berkum et al., 2005; Wicha, Moreno, & Kutas, 2004). Consistent with prediction-by-production accounts, Martin et al. found that the N400 response to articles whose grammatical gender was unexpected was reduced when participants simultaneously performed an articulatory suppression task that taxed the production system (i.e., repeatedly pronouncing a syllable), as compared to two control conditions (i.e., tongue tapping, listening to a recording of one's voice pronouncing the syllable). Thus, limiting the availability of the production system during comprehension appeared to weaken the effects of prediction, implying that comprehenders use their production system to generate predictions.

In this paper, we report a complementary test of prediction-by-production, asking whether increasing the engagement of the production system during comprehension can *enhance* the effects of prediction. In our experiments, participants read sentence contexts that either were (1a) or were not (1b) highly predictive of a final word, and we measured how quickly participants recognised that final word (Experiment 1), named that final word (Experiment 2), or used that word to name a picture (Experiment 3).

(1a) *It was windy enough to fly a… kite.*

(1b) *They went to see the famous… show.*

Crucially, we manipulated the engagement of the production system by asking participants to read the sentence contexts either aloud or silently. The primary purpose of the language production system is to convert a message into sound. According to most theories of spoken production, the speaker constructs the message to be conveyed, activates a network of relevant concepts, accesses lexical items and syntactic information, and then focuses the activation on a single form that provides the input to articulation (Levelt, 1989; see Goldrick, Ferreira, & Miozzo, 2014). Reading aloud engages many of the stages involved in other acts of production, such as the construction of representations of sound (though perhaps not the message since it is provided by the written word itself: Levelt et al., 1999, argued that reading aloud may not require activation of the lexical concept). In particular, it engages many production mechanisms that are not involved in silent reading (where no sound is produced). Thus, reading aloud overall engages the production system to a greater extent than reading silently.

Our method has potential advantages over using articulatory suppression to reduce the availability of the production system during reading. We reasoned that a suppression design would be potentially hard to interpret, because articulatory suppression might interfere with

verbal working memory, and thus disturb the ability of participants to comprehend the sentence contexts and generate predictions. By contrast, reading aloud, like spontaneous speech, demands the use of production processes such as formation of phonetic representations and articulation. Importantly, we do not claim that production processes are not used during silent reading. However, we maintain that their use is necessarily enhanced when reading aloud.

Returning to our experimental design, because the possibility of generating predictions is greater when comprehending highly predictive than less predictive contexts, any effect of prediction on language processing should be greater for (1a) than (1b). And since reading aloud engages production processes to a greater extent than reading silently, the effects of prediction should be further enhanced in the read-aloud mode.

Experiment 1 used a lexical decision task: After reading the sentence contexts, participants were instructed to indicate whether the sentence-final stimulus was a word (*kite*) or a nonword (*kile*). An interaction between predictability (low/high) and reading mode (silent/aloud) would implicate prediction-by-production under conditions when the task relies predominately on comprehension. Experiment 2 used a go/no-go task in which participants were instructed to read aloud the sentence-final stimulus if it was a word, but not if it was a nonword. Here, an interaction would implicate prediction-by-production under conditions when the task involves comprehension (deciding if the stimulus is a word) and also some aspects of production (naming the word). Experiment 3 used a picture naming task in which participants were instructed to name a sentence-final picture stimulus whose name was highly predictable or not given the sentence context. An interaction would implicate prediction-by-production under conditions when the task strongly relies on production (i.e., relies on the whole process of production, from intention to articulation).

2 Experiment 1

2.1 Methods

Our data, materials and commented analysis scripts are available at the Open Science Framework (OSF) website of this project: https://osf.io/xun2v/

2.1.1 Participants

Twenty-four participants, who were Edinburgh University students and native speakers of British English, were paid £6. They had normal or corrected-to-normal vision and reported no language disorders. We set this sample size based on our intuitions about the effectiveness of the manipulation and our experimental design (which used a considerable number of items, 240 per participant).

2.1.2 Design

We used a 2 (Predictability: low vs. high) X 2 (Reading Mode: silent vs. aloud) X 2 (Stimulus Type: nonword vs. word) within-subjects design. Reading Mode was blocked, and the order of reading silently/aloud was a between-subjects manipulation (Order: silent first vs. aloud first). To increase reliance on top-down prediction, contexts and sentence-final stimuli were presented against a white-noise background (we reasoned that the engagement of prediction is particularly likely in the conditions of noise; see Pickering & Gambi, 2018). Moreover, sentence-final stimuli were displayed in a shade of grey that was individually selected to ensure participants' word recognition was impaired but above chance (cf. Stanovich & West, 1979). Prior to the task, we carried out an individual pretest to determine the shade for the sentence-final stimuli.

2.1.3 Materials

To create the stimuli, 24 additional participants were asked to fill in the missing final word for 291 sentences truncated before the last word. From this set, we selected 120 high-predictability sentences (for which the most frequently chosen final word was selected, on average, by 89% [±2] of participants; throughout the paper, the values in square brackets indicate 95% confidence intervals) and 120 low-predictability sentences (for which the most frequently chosen final word was selected, on average, by 20% [±1] of participants). High- and low-predictability sentences were matched for number of words (M_{high} = 7.47 vs. M_{low} = 7.28; $t(238) = 1.20$, $p = .232$). By using the most frequently chosen word (or selecting one in case of a tie), we ensured that all sentences were plausible. To create the nonword stimuli, we replaced the final word of each sentence with a pronounceable nonword matched to that word in length, first letter, and last letter (see Table 1). Each participant saw 240 sentences, 60 per condition, such that participants saw each sentence context once. Trial order was individually randomized. To ensure participants paid attention to sentence contexts in all conditions, forty trials were followed by a simple yes/no comprehension question (e.g., *They found the mouse hiding under the table* followed by *Was the mouse hiding behind the table?*).

Sentence contexts were displayed in a shade of grey identical for all participants (hex #393939). Sentence-final stimuli were presented in a shade of grey that was individually thresholded for each participant, using a pretest. The pretest involved 5 different shades, anchored on the neutral axis in the RGB space and differing in lightness level (#414141, #474747, #4D4D4D, #515151, #565656). Each of the 250 trials began with a central fixation cross displayed against a square of white noise (150 X 150 pixels, black and white; fixation cross shown for 500-1000ms, randomly varied). Then, a single word or a nonword was randomly displayed in one of the shades (300ms). The participant pressed a key to signal whether they saw a word or a nonword (half of trials were words, half nonwords; these stimuli were not used in the main experiment). We identified the shade at which the

participant was closest to 60% accuracy and used it for the sentence-final stimuli in the experiment.

Table 1. An example of the stimuli used in Experiment 1.

	sentence context	word / nonword
high predictability	It was windy enough to fly a	kite / kile
low predictability	They went to see the famous	show / spow

2.1.4 Procedure

Each trial began with a central fixation cross, displayed against a square of white noise, surrounded by a white background (1000-1500ms, randomly varied on trial-by-trial basis). Next, sentence contexts were shown word-by-word (300ms ON, 200ms OFF). On the screen before the final word, the background switched from white to yellow and then the sentence-final stimulus was displayed (300ms). Participants were instructed to read the words on the white background aloud or silently and then to press a key to indicate whether they thought the sentence-final stimulus was a word or a nonword. They were told to never read aloud the sentence-final stimulus. Trials ended after the participant had responded to the stimulus (Figure 1) or after a response to the comprehension question (if a given trial was followed by a comprehension question). The experiment was divided into two blocks of 120 trials (read either aloud or silently), and each block was preceded by 8 practice trials where participants trained to perform the task in the presence of the experimenter. During the main task, if needed, the experimenter reminded the participants about the current reading condition. The experiment lasted about 30 minutes.

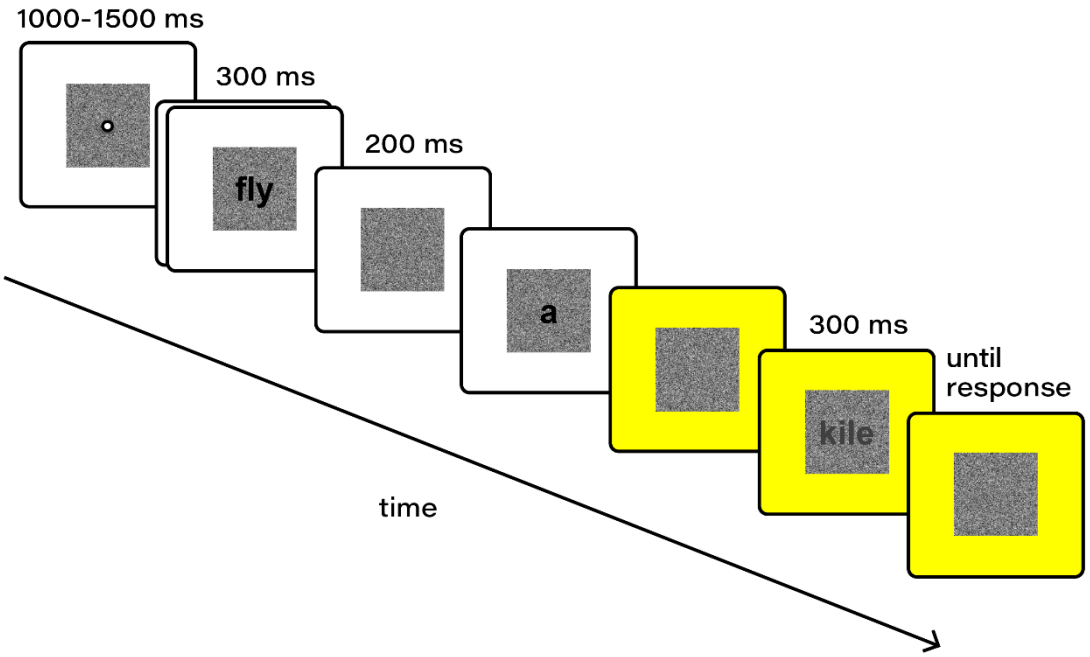

Figure 1. An example of a trial (nonword stimulus) in Experiments 1-2.

2.2 Results

To investigate whether engaging the production system increased the effect of predictability on lexical decisions, we first tested how likely participants were to judge the sentence-final stimulus as a word rather than a nonword. We ran a binomial Generalized Linear Mixed Model (GLMM) with Predictability (low vs. high), Reading Mode (silent vs. aloud), Stimulus Type (nonword vs. word), and Order (silent first vs. aloud first) as fixed effects, and specifying a maximal random structure. Effect-coded contrasts were applied to predictors in all models reported in this paper (Predictability: low was set to -0.5, high to 0.5; Reading Mode: silent was set to -0.5, aloud to 0.5; Stimulus Type: nonword was set to -0.5, word to 0.5; Order: silent first was set to -0.5, high to 0.5). See Supplementary Materials for further details about our models.

Not surprisingly, participants were more likely to respond "word" when reacting to sentence-final stimuli that were real words rather than nonwords, and more likely to respond "word" following high- than low-predictability contexts (*p*s < .001; see means in Table 2 and regression results in Table 3). These effects of Stimulus Type and Predictability marginally interacted ($p = .056$), such that the effect of Predictability was slightly greater for word stimuli than nonword stimuli. However, there was no interaction between Predictability and Reading Mode (i.e., engaging production by reading aloud did not influence the effect of context predictability on the odds of making a "word" response, $p > .250$) and there was no three-way interaction among Predictability, Reading Mode, and Stimulus Type (i.e., engaging production did not influence the effect of predictability on the odds of correctly making a "word" response, $p > .250$). See Supplement for accuracy analyses confirming this pattern of results, and for signal detection (d-prime) analyses of these data.

Next, we analysed key-press reaction times (RT) on trials where participants correctly responded to sentence-final stimuli that were real words. We excluded outliers deviating more than 2.5 SD from each participant's mean (2%), and ran a maximal-structure Linear Mixed Effect (LME) model with Predictability, Reading Mode, and Order as predictors, and with by-subject and by-item random intercepts and slopes. We found that participants were faster to respond following high- than low-predictability contexts, and were overall slower to respond after reading the contexts aloud (*p*s < .001; see means in Table 5 and regression results in Table 6). Although the interaction between Predictability and Reading Mode was not reliable, the descriptive pattern of results was consistent with our expectations: Predictability had a numerically greater effect when participants read the sentence contexts aloud rather than silently ($p = .174$). There were no other effects or interactions. See Supplement for further analyses confirming these findings (i.e., regression models conducted on normalized data and models robust to data contamination).

Finally, we analysed responses to the context comprehension questions, and found that participants were similarly engaged with the task whether they were reading aloud or silently, and whether they were reading high- or low-predictability contexts. They responded correctly to the questions on 90% of trials, and the odds of providing a correct answer did not differ across conditions (*p*s > .250; full results in Supplement).

Table 2. Percentages of "word" responses in Experiments 1 and 2. Table shows mean percentages with 95% confidence intervals by Stimulus Type, Reading Mode, and Predictability.

Experiment 1

	word		nonword	
	low predictability	high predictability	low predictability	high predictability
reading silently	69% [±6]	79% [±5]	28% [±6]	34% [±7]
reading aloud	69% [±6]	82% [±4]	25% [±6]	29% [±7]

Experiment 2

	word		nonword	
	low predictability	high predictability	low predictability	high predictability
reading silently	65% [±6]	83% [±6]	26% [±5]	40% [±6]
reading aloud	72% [±6]	82% [±7]	31% [±6]	51% [±8]

Table 3. GLMM analyses of "word" responses in Experiment 1. Table shows results from the fixed and random effects structure.

Experiment 1

	"word" response

Predictors	*Odds Ratios*	*CI*	*p*
Intercept	1.11	0.88 – 1.40	0.385
Predictability	1.54	1.23 – 1.92	**<0.001**
Reading Mode	0.92	0.78 – 1.09	0.334
Stimulus Type	9.17	6.88 – 12.21	**<0.001**
Order	0.86	0.55 – 1.36	0.530
Predictability * Reading Mode	1.04	0.76 – 1.41	0.816
Predictability * Stimulus Type	1.45	0.99 – 2.12	0.056
Reading Mode * Stimulus Type	1.36	0.94 – 1.96	0.106
Predictability * Order	0.98	0.65 – 1.48	0.930
Reading Mode * Order	0.84	0.61 – 1.16	0.294
Stimulus Type * Order	1.26	0.73 – 2.16	0.411
Predictability * Reading Mode * Stimulus Type	1.26	0.70 – 2.24	0.438
Predictability * Reading Mode * Order	0.71	0.39 – 1.29	0.263
Predictability * Stimulus Type * Order	1.55	0.80 – 3.02	0.194
Reading Mode * Stimulus Type * Order	1.05	0.51 – 2.17	0.893
Predictability * Reading Mode * Stimulus Type * Order	1.14	0.37 – 3.51	0.823
Random Effects			
σ^2			3.29
$\tau_{00\ item}$			0.26
$\tau_{00\ participant}$			0.30
$\tau_{11\ item.stimulus\ type}$			0.06
$\tau_{11\ item.reading\ mode}$			0.10
$\tau_{11\ item.order}$			0.01
$\tau_{11\ item.stimulus\ type:reading\ mode}$			0.22
$\tau_{11\ item.stimulus\ type:order}$			0.13

τ_{11} item.reading mode:order	0.05
τ_{11} item.stimulus type:reading mode:order	0.42
τ_{11} participant.predictability	0.16
τ_{11} participant.reading mode	0.05
τ_{11} participant.stimulus type	0.35
τ_{11} participant.predictability:reading mode	0.14
τ_{11} participant.predictability:stimulus type	0.27
τ_{11} participant.reading mode:stimulus type	0.41
τ_{11} participant.predictability:reading mode:stimulus type	0.34
$N_{participant}$	24
N_{item}	480
Observations	5760
Marginal R^2	0.283

Table 4. GLMM analyses of "word" responses in Experiment 2. Table shows results from the fixed and random effects structure.

Experiment 2

	"word" response		
Predictors	*Odds Ratios*	*CI*	*p*
Intercept	1.47	1.19 – 1.81	**<0.001**
Predictability	2.49	2.10 – 2.95	**<0.001**
Reading Mode	1.46	1.18 – 1.81	**<0.001**
Stimulus Type	7.23	5.19 – 10.08	**<0.001**
Order	1.24	0.82 – 1.87	0.303
Predictability * Reading Mode	1.06	0.81 – 1.39	0.691
Predictability * Stimulus Type	0.93	0.64 – 1.36	0.725

Reading Mode * Stimulus Type	0.88	0.66 – 1.17	0.377
Predictability * Order	0.92	0.70 – 1.19	0.515
Reading Mode * Order	0.96	0.63 – 1.47	0.861
Stimulus Type * Order	1.66	0.89 – 3.11	0.111
Predictability * Reading Mode * Stimulus Type	0.89	0.53 – 1.50	0.667
Predictability * Reading Mode * Order	0.84	0.49 – 1.44	0.532
Predictability * Stimulus Type * Order	0.83	0.45 – 1.53	0.548
Reading Mode * Stimulus Type * Order	0.44	0.25 – 0.77	**0.004**
Predictability * Reading Mode * Stimulus Type * Order	1.78	0.64 – 4.97	0.272

Random Effects

σ^2	3.29
τ_{00} item	0.31
τ_{00} participant	0.32
τ_{11} item.stimulus type	0.28
τ_{11} item.reading mode	0.08
τ_{11} item.order	0.01
τ_{11} item.stimulus type:reading mode	0.14
τ_{11} item.stimulus type:order	0.09
τ_{11} item.reading mode:order	0.08
τ_{11} item.stimulus type:reading mode:order	0.23
τ_{11} participant.predictability	0.04
τ_{11} participant.reading mode	0.27
τ_{11} participant.stimulus type	0.70
τ_{11} participant.predictability:reading mode	0.16
τ_{11} participant.predictability:stimulus type	0.35
τ_{11} participant.reading mode:stimulus type	0.23
τ_{11} participant.predictability:reading mode:stimulus type	0.48

$N_{participant}$	32
N_{item}	480
Observations	7676
Marginal R^2	0.281

Table 5. Participants' RT in Experiments 1, 2, and 3. Table shows means with 95% CI (ms).

Experiment 1		
	low predictability	high predictability
reading silently	879 [±55]	694 [±42]
reading aloud	1211 [±67]	969 [±59]
Experiment 2		
	low predictability	high predictability
reading silently	916 [±29]	867 [±25]
reading aloud	721 [±31]	607 [±34]
Experiment 3		
	low predictability	high predictability
reading silently	597 [±17]	495 [±19]
reading aloud	501 [±21]	321 [±22]

Table 6. LME analyses of RT in Experiment 1. Table shows results from the fixed and random effects structure.

Experiment 1			
		RT	
Predictors	*Estimates ($\hat{\beta}$)*	*CI*	*p*
Intercept	966.97	835.75 – 1098.18	**<0.001**

Predictability	-217.29	-292.94 – -141.64	**<0.001**
Reading Mode	321.22	192.04 – 450.40	**<0.001**
Order	4.31	-256.35 – 264.96	0.974
Predictability * Reading Mode	-75.43	-184.29 – 33.43	0.174
Predictability * Order	-55.95	-194.47 – 82.56	0.429
Reading Mode * Order	256.02	-4.79 – 516.82	0.054
Predictability * Reading Mode * Order	-83.03	-312.04 – 145.98	0.477
Random Effects			
σ^2			290208.62
$\tau_{00\ item}$			14509.61
$\tau_{00\ participant}$			102570.50
$\tau_{11\ item.reading\ mode}$			6314.52
$\tau_{11\ item.order}$			1067.97
$\tau_{11\ item.reading\ mode:order}$			100415.47
$\tau_{11\ participant.predictability}$			15785.02
$\tau_{11\ participant.reading\ mode}$			89507.28
$\tau_{11\ participant.predictability.reading\ mode}$			15139.75
$N_{participant}$	24		
N_{item}	240		
Observations	2122		
Marginal R^2	0.122		

Table 7. LME analyses of RT in Experiment 2. Table shows results from the fixed and random effects structure.

Experiment 2
RT

Predictors	*Estimates ($\hat{\beta}$)*	*CI*	*p*
Intercept	774.06	747.01 – 801.10	**<0.001**
Predictability	-42.49	-60.11 – -24.87	**<0.001**
Reading Mode	-114.54	-138.98 – -90.11	**<0.001**
Order	-36.13	-63.74 – -8.53	**0.010**
Predictability * Reading Mode	-15.98	-31.96 – -0.00	0.050
Predictability * Order	-1.24	-19.72 – 17.23	0.895
Reading Mode * Order	6.21	-17.98 – 30.41	0.615
Predictability * Reading Mode * Order	-7.47	-23.07 – 8.13	0.348
Random Effects			
σ^2			117866.14
$\tau_{00\ item}$			1905.83
$\tau_{00\ participant}$			3647.01
$\tau_{11\ item.reading\ mode}$			1495.36
$\tau_{11\ item.order}$			3858.17
$\tau_{11\ item.reading\ mode:order}$			733.35
$\tau_{11\ participant.predictability}$			508.95
$\tau_{11\ participant.reading\ mode}$			2687.03
$\tau_{11\ participant.predictability.reading\ mode}$			140.24
$N_{participant}$	29		
N_{item}	240		
Observations	2339		
Marginal R^2	0.119		

Table 8. LME analyses of RT in Experiment 3. Table shows results from the fixed and random effects structure.

Experiment 3

		RT	
Predictors	*Estimates ($\hat{\beta}$)*	*CI*	*p*
Intercept	490.15	441.58 – 538.71	**<0.001**
Predictability	-139.30	-192.74 – -85.85	**<0.001**
Reading Mode	-133.49	-170.23 – -96.75	**<0.001**
Order	1.81	-83.93 – 87.55	0.967
Predictability * Reading Mode	-69.69	-106.60 – -32.77	**<0.001**
Predictability * Order	41.63	-14.00 – 97.26	0.142
Reading Mode * Order	77.10	3.13 – 151.06	**0.041**
Predictability * Reading Mode * Order	38.15	-37.64 – 113.95	0.324
Random Effects			
σ^2			43000.64
$\tau_{00\ item}$			16659.20
$\tau_{00\ participant}$			14213.25
$\tau_{11\ item.reading\ mode}$			412.44
$\tau_{11\ item.order}$			1658.97
$\tau_{11\ item.reading\ mode:order}$			4143.30
$\tau_{11\ participant.predictability}$			3874.15
$\tau_{11\ participant.reading\ mode}$			8724.16
$\tau_{11\ participant.predictability.reading\ mode}$			2726.94
$N_{participant}$	31		
N_{item}	120		
Observations	3025		
Marginal R^2	0.176		

3 Experiment 2

Experiment 1 did not find statistical evidence that engaging the production system enhanced the use of prediction in a lexical decision task, as indicated by the lack of a significant interaction between Predictability and Reading Mode. However, the descriptive pattern of participants' response times in Experiment 1 was suggestive of the possibility that production does affect prediction (i.e., predictability had a larger numerical effect on response times when reading aloud versus silently). At this stage, we speculated that an effect of production on prediction might be easier to identify when the experimental task itself relies on the production system (see Huettig, 2015). Thus, in Experiment 2 we investigated whether engaging the production system causes larger effects of predictability in a task that required a spoken response.

As before, we asked participants to read high- and low-predictability sentences either aloud or silently, but now they also had to read aloud the final word if they decided it was a word, and not otherwise. In this experiment, participants' decisions whether to name the sentence-final stimulus reflect their ability to comprehend, as in Experiment 1, but the time to articulate that name reflects processes of spoken language production. If prediction-by-production helps subsequent production processes, then participants should take less time to articulate the name after reading high- than low-predictability contexts, and this boost should increase when they read the contexts aloud.

3.1 Methods

3.1.1 Participants

We recruited 32 further participants from the same population and on the same terms as in Experiment 1. We increased the sample size based on the possibility that Experiment 1

was underpowered (see Supplement for an estimation of the sample size needed to detect prediction-by-production effects).

3.1.2 Materials, procedure, and design

The experiment was identical to Experiment 1, except that participants were instructed to read the sentence-final stimulus out loud if it was a word and not do so otherwise. Responses were recorded for 3000ms from the onset of the sentence-final stimulus, using a microphone positioned in front of the participant.

3.2 Results

Two trained coders analysed the recordings from the experiment. For each trial, they identified whether participants named the sentence-final stimulus, and calculated naming time (i.e., time from stimulus onset until participants started articulating the name). The inter-rater reliability between coders for this measure was in the excellent range (two-way random, consistency ICC = .99; calculated on 6% data; Cicchetti, 1994).

To test whether prediction-by-production affected the decisions to read the sentence-final stimuli, we analysed the odds of naming each sentence-final stimulus across conditions, using a binomial GLMM with Predictability, Reading Mode, Stimulus Type, and Order as fixed effects, and including a maximal random structure. The means are presented in Table 2 and the results from the model in Table 4.

We found that participants were more likely to name the sentence-final stimulus after a high-predictability context (an effect of Predictability, $p < .001$), and were more likely to name the stimulus if it was a word (an effect of Stimulus Type, $p < .001$). However, these effects showed no tendency toward an interaction ($p > .250$), and there were no further interactions involving both Predictability and Reading Mode, suggesting that the effects of predictability were similar whether participants were reading aloud or silently ($ps \geq .250$).

We also found that participants were more likely to name the sentence-final stimulus after they had read the context sentence aloud than silently (an effect of Reading Mode, $p <$ .001), perhaps because it was more difficult for them to inhibit their spoken response in this case. Furthermore, the effect of Reading Mode interacted with Order and Stimulus Type ($p =$ .004), reflecting the fact that participants who read aloud in the second block made more "word" responses (i.e., meaning they read the final stimuli out loud) after reading aloud than silently both for word and nonword stimuli (word stimuli: M_{aloud} = 77% [±7], $M_{silently}$ = 68% [±7]; nonword stimuli: M_{aloud} = 41% [±7], $M_{silently}$ = 35% [±6]), whereas participants who read aloud in the first block did so for nonword stimuli but not for word stimuli (word stimuli: M_{aloud} = 78% [±7], $M_{silently}$ = 79% [±5]; nonword stimuli: M_{aloud} = 41% [±8], $M_{silently}$ = 31% [±6]). However, this finding is not informative of linguistic prediction (the statistical interaction did not include Predictability), and can in fact be attributed to practice effects: The proportion of correct "word" responses increased across blocks, so that for participants who started off reading silently it was higher in reading aloud, and for participants who started off reading aloud it was higher in reading silently. Consistently, additional signal detection analyses showed that d-prime sensitivity increased across blocks (see Supplement).

Next, we analysed participants' naming times for the sentence-final stimuli. When the final stimulus was a word, participants named it on 77% [±2] of trials. From these trials, we excluded by-participant outliers as per Experiment 1 (<1%) and then trials where in reading aloud participants were still reading the context after the onset of the sentence-final stimulus (a further 18%). Because of this, three participants lost more than half observations in read-aloud mode and were excluded from analysis (leaving n = 29). We ran a maximal-structure LME model with naming time as the dependent variable, Predictability, Reading Mode, and

Order as fixed effects, and with by-subjects, by-items random intercepts and slopes. The means are presented in Table 5 and the regression results in Table 7.

Unsurprisingly, we found that naming times were affected by Predictability and Reading Mode: Participants were faster to name the final word after a high-predictability context, and were also faster after they had read the context aloud (both *p*s < .001). Crucially, and just as in Experiment 1, we observed that the numerical difference between high- and low-predictability contexts was greater in reading aloud. The interaction between Predictability and Reading Mode was marginal in our main LME model (p = .050), but significant in further analyses accounting for the non-normal distribution of our response time data (i.e., robust models and analyses on normalized data; see Supplement). In addition, we found an effect of Order (p = .010): Participants who read aloud in the first block were faster to name the final word than those who read aloud in the second block (M_{first} = 758ms [±20], M_{second} = 822ms [±23]), perhaps reflecting fatigue effects.

When answering the comprehension questions, participants were more likely to produce a correct response following highly predictable contexts (M_{high} = 86% vs. M_{low} = 77%; p = .024), and marginally more likely to answer correctly in reading aloud (M_{aloud} = 83% vs. $M_{silently}$ = 80%; p = .055), but these effects did not interact (*p*s = .594; full results in Supplement). Thus there is no reason to believe that these minor differences affected the key interaction between Predictability and Reading Mode when participants named the final word. Overall, participants responded correctly on 82% of trials.

4 Experiment 3

Experiment 2 revealed a numerical pattern consistent with the claim that the production system plays a role in generating predictions: The effect of contextual predictability appeared to be enhanced when the production system had been engaged during

context processing. However, the critical interaction was marginal in our main analyses (though significant in the supplementary analyses).

We conjectured that prediction-by-production effects might be clearest when the task engages all processes of spoken production. An important demonstration of production being implicated in prediction comes from the ultrasound imaging study of Drake and Corley (2015), where context predictability affected the tongue movements performed while naming pictures. Thus, in Experiment 3, we replaced word naming with picture naming.

4.1 Methods

4.1.1 Participants

We recruited 32 further participants from the same population and on the same terms as in the previous experiments. We set sample size based on Experiment 2.

4.1.2 Design

We used a 2 (Predictability: low vs. high) X 2 (Reading Mode: silent vs. aloud) within-subjects design. Participants read high- and low-predictability sentence contexts, half of the time aloud, and half of the time silently (the order of reading silently/aloud was counter-balanced between participants; Order: silent first vs. aloud first). The sentence contexts were followed by a picture stimulus which participants named aloud into a microphone.

4.1.3 Materials

We developed new stimuli appropriate for picture-naming. There were 60 high-predictability and 60 low-predictability sentence contexts (length-matched: M_{high} = 8.08 vs. M_{low} = 7.81; $t(118) = -1.17$, $p = .244$). Predictability values were determined by 24 additional

participants, who filled in the final word of 206 different contexts. The most frequent continuation was used, on average, by 92% [±2] of participants for high-predictability contexts, and 16% [±1] for low-predictability contexts. Sentence contexts were displayed in a shade of grey identical for all participants (#393939). Each high-predictability context was paired with a picture whose name was the most frequent continuation of the context, and each low-predictability context with a picture whose name was a plausible, but not the most frequent, continuation of the context (see Table 9). Pictures could be named with a single word and had high name agreement (M = .89) and high name frequency (M = 4.16). The pictures used for high- and low-predictability contexts were matched for name frequency (M_{high} = 4.22 vs. M_{low} = 4.09, $t(116) = 1.41$, $p = .161$) and name length in syllables (M_{high} = 1.53 vs. M_{low} = 1.73, $t(116) = 1.39$, $p = .165$).[1]

Each participant saw two lists each comprising 30 high- and 30 low-predictability sentences. Lists were matched for context length, picture name agreement, name frequency, and name length in syllables (all $ps \geq .108$). Trial order was randomized within each list for each participant. The order of lists was counterbalanced between participants. Twelve trials were followed with a yes/no context comprehension question. Pictures and norms for picture name agreement were taken from the Bank of Standardized Stimuli (BOSS v.2; Brodeur, Guérard, & Bouras, 2014). Norms for picture name frequency and length were taken from SUBTLEX-UK (van Heuven, Mandera, Keuleers, & Brysbaert, 2014).

---

[1] Note that in Experiments 1-2 we did not control for such differences at the stage of stimuli selection. Instead, we controlled for them in our statistical analyses – we re-ran our main models controlling for these variables and found that adding them did not affect the pattern of results. The description of these analyses can be found in the Supplement.

Table 9. An example of the stimuli used in Experiment 3.

	sentence context	picture
high predictability	*It was windy enough to fly a*	
low predictability	*They went to see the famous*	

4.1.4 Procedure

Trials began with a fixation cross, followed by a sentence context presented as in Experiments 1-2. Each context was followed by a picture and participants were instructed to name it with a single word, as fast as possible. Responses were recorded for 3000ms from picture onset by a microphone positioned in front of the participant. Trials ended after the timeout of the recording (Figure 2) or the question response (if a given trial was followed by a comprehension question).

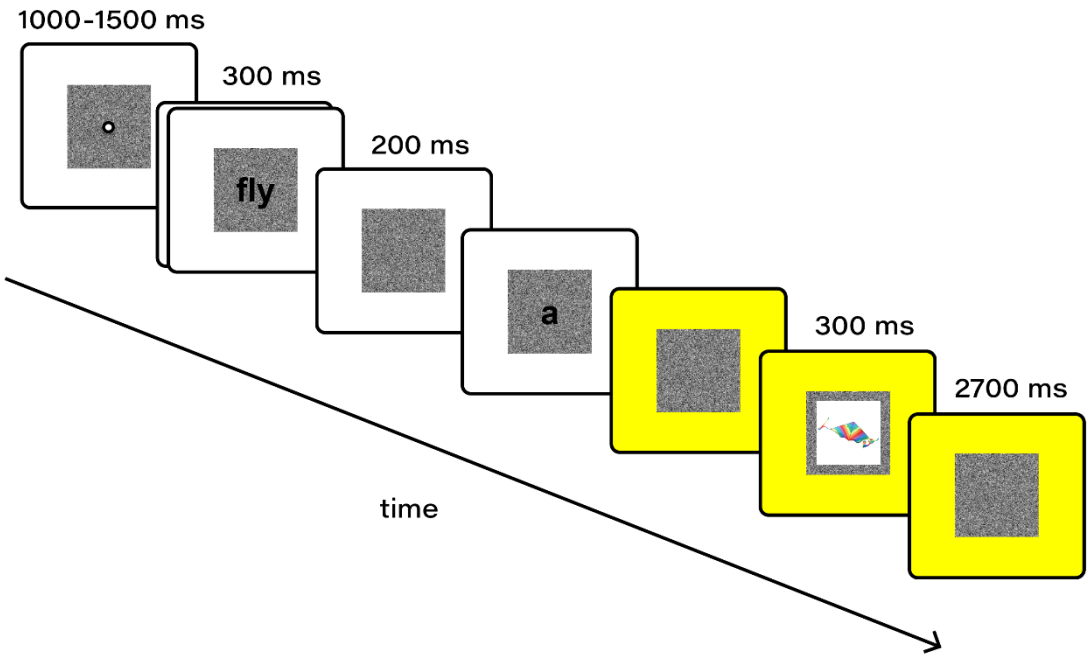

Figure 2. An example of a trial in Experiment 3.

4.2 Results

Three trained coders calculated participants' RT (i.e., time from picture onset until naming onset). The inter-rater reliability was again excellent (two-way random, consistency ICC = .99; calculated on 6% data). Prior to the analyses we removed by-participant outliers as per Experiments 1-2 (3%), and trials on which participants were still reading the sentence context aloud after the onset of the picture (16%). As a result of this, one participant lost more than half observations in read-aloud mode. We excluded their data from further analyses (leaving n = 31).

To test whether reading the context aloud enhanced the facilitative effect of prediction on picture naming, we ran a maximal-structure LME model with Predictability, Reading Mode, and Order as fixed effects, and by-subjects, by-items random intercepts and slopes. As

in Experiment 2, participants named the pictures faster after a high-predictability context, and also faster after reading the sentence context aloud (both *p*s < .001; see Tables 5 and 8). But most importantly, we now observed a reliable interaction between Predictability and Reading Mode: The effect of Predictability on naming times was greater when participants had read the sentence contexts aloud than silently ($p < .001$). Once again, these findings were confirmed by further analyses (i.e., robust models and analyses on normalized data; see Supplement).

There was also an additional interaction between Reading Mode and Order ($p = .041$), such that the effect of reading aloud tended to be greater in participants who read aloud in the second block (M_{aloud} = 398ms [±24], $M_{silently}$ = 564ms [±20]) than those who read aloud in the first block (M_{aloud} = 425ms [±22], $M_{silently}$ = 529ms [±16]). A plausible explanation for this finding is that experience with the task amplified the simple facilitative effect of reading aloud on naming speed. But importantly, these effects did not interact further with Predictability ($p > .250$), providing no evidence that linguistic prediction was affected by whether participants first read aloud or silently.

Finally, we found no evidence that the key interaction between Predictability and Reading Mode could be explained by general context comprehension: Although participants were more likely to correctly answer comprehension questions after high- than low-predictability contexts (M_{high} = 90%, M_{low} = 78%; $p = .012$), there was no effect of Reading Mode, or interactions with Reading Mode (*p*s > .250; see Supplement). Participants responded correctly on average on 84% of trials.

5 Mini Meta-Analysis

Three experiments tested whether engaging the production system would enhance the effects of linguistic prediction on three different tasks – lexical decision, word naming, and

picture naming. In analysing the response times of Experiment 3, which used a picture naming task, we observed a robust interaction between Predictability and Reading Mode, indicating that Predictability had a greater effect on response times when participants read the sentence contexts aloud rather than silently. However, in Experiment 1 (lexical decision) the relevant interaction was not statistically significant, and in Experiment 2 (word naming), it was marginally significant.

An important question is whether this pattern of findings reflects a theoretically meaningful difference among those studies. In particular, whereas the task used in Experiment 1 relied on production processes to a very small extent (i.e., manual lexical decision), Experiment 2 relied on production more strongly (i.e., spoken go/no-go), and Experiment 3 involved a measure that very strongly engaged production processes (i.e., picture naming involves production from intention to articulation), which raises the possibility that production affects use of prediction only in tasks that themselves strongly rely on the production system. Alternatively, however, the difference in statistical significance between the experiments could potentially reflect lack of power in Experiments 1-2, which used a smaller number of participants than Experiment 3, or perhaps result from chance variation in the observed statistical effect. Consistent with this, the numerical pattern of response times in Experiments 1-2 did follow the direction expected if an interaction had been present.

To choose between these alternatives we conducted a so-called mini meta-analyses of our three experiments (Cumming, 2014; Cumming, Fidler, Kalinowski, & Lai, 2012; Goh, Hall, & Rosenthal, 2016). For each experiment, we calculated Cohen's *d* for the size of the Predictability and Reading Mode interaction (Table 10), and fit random-effects meta-analytic models using the package metafor v.2.0-0 (Viechtbauer, 2010). To test if the overall effect size across the three experiments was greater than zero, we fit a model that only contained an

intercept term; that intercept term was statistically significant (B = -.51 (.26), CI = [-1.02, -.04], z = -1.97, p = .048), suggesting that the summary effect size of the interaction across experiments was -0.51. To test whether task type (how strongly it relied on production) moderated the interaction effect size, we conducted a second regression in which that distinction was included as a moderator (i.e., we compared the effect size among the three experiments). However, the effect of task type was not statistically significant (B = -.23 (.32), CI = [-.86, .39], z = -.73, p > .250). These data provide no support for the claim that prediction-by-production is limited to contexts that strongly rely on the production system. But crucially, our results indicate that the critical effect of production on prediction is robust when considering the totality of our data.

We also used a mini meta-analysis to conduct a control analysis on the possibility that participants engaged more deeply with the experimental materials when reading them aloud. Recall that participants in our experiments answered comprehension questions. In Experiments 1 and 3, accuracy on those questions was equivalent across the two reading modes but, in Experiment 2, accuracy was slightly higher in the reading aloud condition – a 3% difference that was marginally significant. We conducted a random effects meta-analysis across Cohen's d effect sizes calculated for each experiment (difference in proportion of correct answers between conditions). Across experiments, we found that the estimated effect size was small and not significantly different from zero (B = .23 (.25), CI = [-.25, .72], z = .95, p > .250). We thus conclude that enhanced context comprehension in the reading aloud condition is unlikely to explain the critical interaction between reading mode and predictability found in our studies.

6 'Omnibus' LME Analysis

To further verify the findings from our mini meta-analysis, we combined the data from all three experiments (i.e., we merged the final datasets involved in the LME analyses

reported above) and ran a LME model with Predictability (low vs. high), Reading Mode (silent vs. aloud), and Task (experiment 1 vs. experiment 2 vs. experiment 3) as predictors, and by-participant and by-item random intercepts and slopes (effect-coded contrasts were applied to Predictability and Reading Mode in the same manner as in the LME analyses reported above; the reference level of Task was set to experiment 1).

The results are presented in Table 11. There were main effects of Predictability ($p < .001$) and Reading Mode ($p < .001$), indicating that response times were longer for low than high predictability contexts and for reading aloud than silently. More importantly, the critical interaction between Predictability and Reading Mode was significant ($p = .022$), but the three-way interaction between Predictability, Reading Mode, and Task was not ($ps > .281$). These findings are consistent with the meta-analysis: They imply that prediction is facilitated by engaging production, but they provide no evidence for the possibility that such facilitation is stronger in tasks that more strongly rely on the production system.

In addition, the analysis revealed some effects of Task: There was a significant two-way interaction between Task and Reading Mode ($ps < .001$), such that in Experiment 1 reading aloud was associated with longer, and in Experiments 2-3 with shorter response times compared to reading silently (see Table 5). This difference might have been caused by the fact that in the reading aloud condition of Experiment 1 participants had to switch between the task of reading out the sentence contexts and producing a manual response to the sentence-final stimuli. In Experiments 2-3 the response was spoken and so there was no task-switching. Moreover, the analysis also found a significant and unexpected interaction between Task and Predictability ($ps < .032$), indicating that the facilitative effect of prediction was greater in Experiment 1 than in Experiments 2-3 (Table 5); it could be that naming – as an overlearned task – is less affected by contextual support than a less-practiced lexical decision task.

Table 10. Cohen's *d* and 95% CI for the size of the Predictability and Reading Mode interaction in Experiments 1-3, including information about the experimental task, sample size (after exclusions), and average number of items per participant per design cell (after exclusions).

	task	sample	items/participant/cell	d
Experiment 1	button-press	24	22	-0.31 [±0.57]
Experiment 2	spoken go/no-go	29	21	-0.45 [±0.52]
Experiment 3	picture naming	31	24	-0.78 [±0.53]

Table 11. An "omnibus" LME analysis of RT data from Experiments 1-3. Table shows results from the fixed and random effects structure.

"Omnibus" LME analysis of data from Experiments 1-3

		RT	
Predictors	*Estimates ($\hat{\beta}$)*	*CI*	*p*
Intercept	973.86	890.67 – 1057.04	**<0.001**
Predictability	-221.57	-273.25 – -169.89	**<0.001**
Reading Mode	324.58	239.26 – 409.89	**<0.001**
Task: Exp. 2	-489.86	-601.13 – -378.60	**<0.001**
Task: Exp. 3	-492.57	-603.70 – -381.45	**<0.001**
Predictability * Reading Mode	-82.12	-152.60 – -11.65	**0.022**
Predictability * Task: Exp. 2	99.93	38.06 – 161.80	**0.002**
Predictability * Task: Exp. 3	78.73	7.24 – 150.22	**0.031**
Reading Mode * Task: Exp. 2	-679.68	-795.52 – -563.84	**<0.001**
Reading Mode * Task: Exp. 3	-455.85	-569.39 – -342.30	**<0.001**

Predictability * Reading Mode * Task: Exp. 2	53.48	-44.00 – 150.96	0.282
Predictability * Reading Mode * Task: Exp. 3	13.22	-79.71 – 106.16	0.780

Random Effects

σ^2	122213.55
$\tau_{00\ item}$	11018.18
$\tau_{00\ participant}$	40727.45
$\tau_{11\ item.reading\ mode}$	850.18
$\tau_{11\ participant.predictability}$	6769.62
$\tau_{11\ participant.reading\ mode}$	39482.02
$\tau_{11\ participant.predictability.reading\ mode}$	7153.42
$N_{participant}$	84
N_{item}	337
Observations	7425
Marginal R^2	0.276

7 General Discussion

We investigated how activating the production system by reading aloud potentially enhances the effects of linguistic prediction on language processing. Across three experiments, we found that the effect of contextual predictability on response time was greater when participants had read the sentence contexts aloud rather than silently: This effect was clearest in a picture naming task that itself strongly relied on production processes (Experiment 3), was statistically marginal in a go/no-go task that engaged production to a lesser degree (Experiment 2), and was not significant in a manual lexical decision task (Experiment 1). Although the evidence from Experiments 1 and 2 is not strong when considered in isolation, two analyses involving combined data from all our studies – one a mini meta-analysis and one an omnibus mixed model analysis – suggested that the overall evidence was consistent with a signficant effect of reading aloud on prediction.

The importance of these studies is that they clarify the causal relationship between engagement of the production system and use of predictive processing in language. By contrasting reading silently versus reading aloud, we compared the effects of prediction in situations where the production system was fully available but engaged in comprehension to a greater or lesser extent. We found that effects of linguistic prediction were enhanced when the engagement of the production system was greater, and that this degree of enhancement was sufficient to significantly affect behavioural responses (at least in the picture naming task of Experiment 3). This finding accords with prior work that suggested a link between prediction and language production (e.g., Drake & Corley, 2015; Federmeier, Kutas, & Schul, 2010; Mani & Huettig, 2012; Rommers et al., 2020), and with the ERP study by Martin et al. (2018) that showed linguistic predictions can be reduced by supressing the production system. Critically, our study goes beyond that work by demonstrating that the production system can facilitate linguistic predictions, and that this facilitation can affect behavioural, rather than just neural, responses. Our behavioural evidence is particularly important because there is controversy about the reliability of some supposed neural signatures of prediction (Nieuwland et al., 2019). By contrast, there is no controversy about the finding that prediction facilitates behavioural responses such as the ones studied here, and so our data can be more directly interpreted.

We now interpret our findings in relation to Pickering and Garrod's (2013) account of how the language production system is involved in prediction. Pickering and Garrod argued that comprehenders use a particular form of prediction-by-production that they call prediction-by-simulation. When people produce utterances, they learn the relationship between their intention (or production command) and the linguistic (and non-linguistic) properties of their intended utterance, such as the sounds of the words that they utter. Over time, they learn to predict aspects of their experience of producing an utterance, as soon as

they have the intention to produce that utterance, using so-called *forward models* (cf. Hickok, 2012). For example, they might develop the intention to say *kite*, and then rapidly predict that they will experience themselves saying /kaIt/ (or perhaps just the initial /k/). When they hear someone else speaking, they covertly imitate that person and (making allowances for differences between that person and themselves) use their forward models to predict their upcoming experience of what the speaker will say next. So if they covertly imitate someone saying *It was windy enough to fly a…*, they then predict the experience of hearing /kaIt/ (or /k/).

Our evidence that context predictability impacts sentence processing is compatible with several prediction accounts (e.g., Huettig, 2015; Kutas & Federmeier, 2011; Kuperberg & Jaeger, 2016; Pickering & Gambi, 2018). But our finding that reading aloud enhances prediction provides support for Pickering and Garrod's (2013) account. Reading aloud increases the engagement of the production system – and, since the production system is used in prediction, reading aloud increases prediction. In this sense, our data are also compatible with other prediction-by-production models, notably the early computational model of Chang, Dell, and Bock (2006) which proposed that prediction in comprehension is carried out by the production system, and Dell and Chang's (2014) framework which provides computational evidence for equating prediction and production.

One interesting aspect of our data is that reading aloud led to stronger predictions (i.e., shorter naming times), but not necessarily more accurate ones (i.e., the proportion of correct "word" responses did not improve in reading aloud). This finding is again consistent with Pickering and Garrod (2013) who propose that comprehenders use their production system to simulate how the perceived utterance will unfold. It is thus possible that engaging the production particularly strongly makes the prediction mechanism more "selfish", leading

to predictions that more closely reflect what would the comprehender say, rather than what is the likely continuation from the speaker.

In addition, this pattern of results resonates with the claim that linguistic prediction involves production up to the stage of phonological planning (Pickering & Garrod, 2013; Pickering and Gambi, 2018; for a discussion, Huettig, 2015). The phonological stage of language production is typically associated with a small number of planned utterances (Levelt et al., 1999; Peterson & Savoy, 1998). For example, whereas several representations may retain a high activation status at the lexico-semantic stage (*It was windy enough to fly a…* PLANE, BALLOON, KITE), at the phonological stage a single representation is likely to remain active (/kaIt/). Since overt production necessarily involves phonological planning to a greater extent than silent comprehension, our reading aloud manipulation might have increased the role of phonology in the forward model, in turn constraining the number of generated predictions. Moreover, fewer predictions to choose from should mean smoother response selection, particularly when participants produce a spoken response (i.e., shorter naming times in Experiment 3).

It is important to note that there has been some controversy about phonological prediction, with the findings of Delong et al. (2005) not being replicated by Nieuwland et al. (2018). However, there is clear evidence of phonological prediction in another study using ERPs (Ito, Gambi, Pickering, Fuellenbach, & Husband, 2020) and in the "visual world" eye-tracking paradigm (Ito, Pickering, & Corley, 2018; Kukona, 2020). It thus appears most likely that comprehenders do engage in phonological prediction, but prediction of phonology is more limited than semantic or other types of prediction, presumably because it implicates later stages of the production process (see Pickering & Gambi, 2018). In our study, reading aloud engages phonological processing, and therefore is particularly good at enhancing phonological predictions. However, it may be that reading aloud also enhances other aspects

of prediction (such as semantics) – we cannot distinguish between these alternatives based on our results.

This brings us to one of the main questions related to linguistic prediction: under what circumstances can comprehenders use the production system to predict? Although this issue remains under debate, some preliminary conclusions may be drawn at this stage. In particular, it appears that comprehenders might not always rely on prediction-by-production. There is some indication that this type of prediction may be resource-intensive (Hintz, Meyer, & Huettig, 2017; Ito, Corley, Pickering, Martin, & Nieuwland, 2016) and so its use could be limited when cognitive resources are scarce (cf. Huettig & Janse, 2016; Ito, Corley, & Pickering, 2017). More importantly, there is evidence that prediction can be achieved without the contribution of forward models (i.e., comprehenders might predict based on the activation of event schemes; Amsel, DeLong, & Kutas, 2015; Kukona et al., 2011; Metusalem et al., 2012). Indeed, it seems that prediction may sometimes occur without engaging the late production processes (phonology, articulation), and there may be routes to prediction that do not involve production at all (Huettig, 2015; Pickering & Gambi, 2018). On such occasions, comprehenders might instead rely on prediction-by-association (i.e., they might predict based on the spreading activation between representations; Collins & Loftus, 1975).

However, there are times when the use of prediction-by-production seems likely. For example, prediction-by-production could be particularly engaged when the spoken or written signal is distorted by environmental noise. Nutall et al. (2016) observed that TMS-elicited motor evoked potentials in the cortical lip area were larger when comprehenders listened to distorted rather than normal speech, particularly when they listened to lip-articulated sounds. Further, Adank, Hagoort, and Bekkering (2010) showed that comprehension of accented speech in noise can be improved by training the production system – participants whose training involved imitating the speaker's accent were better at subsequently understanding

this accent than control participants who did not imitate. Although these studies do not directly show that such improvements in comprehension are due to prediction-by-production, they are compatible with the idea that this mechanism may be used to support language processing in the conditions of increased difficulty.

The use of prediction-by-production could also be encouraged in contexts that involve an overall heightened activation of the production system. Specifically, it has been argued that comprehenders might be inclined to use this mechanism in dialogue (Pickering & Garrod, 2021; Scott, McGettigan, & Eisner, 2009). Due to its collaborative nature, dialogue requires the interlocutors to carefully coordinate each other's utterances (Clark & Wilkes-Gibbs, 1986; Clark, 1996). The ability to predict how the current speaker will continue would prove very useful. In addition, dialogue requires a near-constant activation of the production system – even while listening, the comprehenders provide feedback to the speaker, for instance by expressing understanding or signalling the desire to take the speaking turn.

Indeed, there is some evidence that prediction-by-production could be more prominent in contexts that strongly rely on production. In Hintz and Meyer (2015), participants listened to basic mathematical equations and made fixations to numbers corresponding to the results of these equations. Fixation latencies were shorter on trials where participants heard an incomplete equation (*one plus five is*) and spoke the expected result (*six*), compared to trials where they heard a complete equation and did not speak (*one plus five is six*), consistent with the notion that prediction is facilitated in the conditions of an increased involvement of production. However, since this study did not manipulate predictability (i.e., due to the simplicity of the equations the results were always highly predictable), it may be that the faster fixations were instead caused by participants being more engaged with the task in the speaking condition (this interpretation is supported by the finding that other, non-predictive fixations were also faster in this condition). Clearer

evidence comes from Hintz et al. (2016) where sentence completions were read more quickly after predictable than unpredictable contexts. Importantly, such an advantage occured only in a task where participants also named pictures that appeared as part of the completions, implying that behavioural effects of prediction might occur especially in contexts where the production system is overall highly activated.

At the first glance, it may seem that our data conform with this hypothesis. The effect of prediction-by-production appears to be particularly well pronounced in the task that more strongly relied on the production system (i.e., picture naming), compared to the tasks that engaged production to a lesser extent (i.e., go/no-go and manual lexical decision). However, the statistical analysis of our data does not support these intuitions – we found no evidence for an effect of experimental task on prediction-by-production (as attested both by the mini meta-analysis and the LME analysis on combined data). Our study therefore provides little insight into any relationship between the overall activation of the production system and the use of prediction-by-production.

One interpretational issue for our study is whether activating the production system enhanced prediction in an indirect fashion by facilitating the comprehension of the context. For example, one could speculate that reading aloud might have been more engaging for our participants, and thus encouraged them to process the contexts more deeply (see Hintz & Meyer, 2015). However, to our knowledge, no existing theory makes this claim. In fact, a converse argument is that allocating resources to overt production might hinder deep processing of the text, particularly in fast reading (recall the quick presentation rate in our paradigm). Importantly, our control analysis showed that comprehension of context sentences was equivalent across the different conditions of reading aloud versus silently (note that Martin et al., 2018 reported a similar result). Thus, it seems unlikely that context comprehension or, more generally, the allocation of mental resources, differed greatly

between reading aloud and silently. That said, our measure of comprehension has limitations – the comprehension questions appeared after the participant had read the context sentence and responded to the sentence-final stimuli; this is a common procedure used to avoid distracting participants from the main task (e.g., Hintz et al., 2016; Martin et al., 2018). The consequence of this procedure is that our measure taps into off-line, rather than on-line comprehension. For this reason, our analysis of the comprehension questions cannot rule out the possibility that reading aloud enhanced on-line processing to some degree.

In sum, our study provides some evidence that the production system plays a role in making linguistic predictions, and that such prediction-by-production can be observed directly at the behavioural level. These findings contribute to the growing body of research suggesting that production may be causally implicated in prediction, and set the stage for a more precise delineation of the relationship between production, comprehension, and prediction.

Author Contributions

All authors contributed to the development of the study concept and design. Testing, data collection and analysis were performed by J. R. L. The manuscript was drafted by J. R. L. and M. J. P. and H. R. provided critical revisions. All authors approved the final version of the manuscript for submission. We would like to thank Ivana Bačanek for help with developing Figures 1 and 2.

Acknowledgements

This research was supported by a School Research Support Grant from the School of Philosophy, Psychology and Language Sciences, University of Edinburgh to J. R. L., and by a grant from the ESRC (ES/L01064X/1) to H.R.

References

Adank, P., Hagoort, P., & Bekkering, H. (2010). Imitation improves language comprehension. *Psychological Science*, *21*(12), 1903-1909.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*, 247–264.

Amsel, B. D., DeLong, K. A., & Kutas, M. (2015). Close, but no garlic: Perceptuomotor and event knowledge activation during language comprehension. *Journal of Memory and Language, 82*, 118-132.

Brodeur, M. B., Guérard, K., & Bouras, M. (2014). Bank of Standardized Stimuli (BOSS) phase II: 930 new normative photos. *PLoS ONE, 9*, e106953.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological review, 113*(2), 234-272.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290.

Clark, H. H. (1996). *Using language.* Cambridge university press.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*(1), 1-39.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review, 82*(6), 407-428.

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, *25*(1), 7-29.

Cumming, G., Fidler, F., Kalinowski, P., & Lai, J. (2012). The statistical recommendations of the American Psychological Association Publication Manual: Effect sizes, confidence intervals, and meta-analysis. *Australian Journal of Psychology*, *64*(3), 138-146.

D'Ausilio, A., Jarmolowska, J., Busan, P., Bufalari, I., & Craighero, L. (2011). Tongue corticospinal modulation during attended verbal stimuli: Priming and coarticulation effects. *Neuropsychologia*, *49*(13), 3670-3676.

Dell, G. S. & Chang, F. (2014). The P-Chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B, 369*: 20120394, 1471-2970.

Drake, E., & Corley, M. (2015). Articulatory imaging implicates prediction during spoken language comprehension. *Memory & Cognition, 43,* 1136-1147

Farmer, T. A., Brown, M., & Tanenhaus, M. K. (2013). Prediction, explanation, and the role of generative models in language processing. *Behavioral and Brain Sciences, 36*(3), 211-212.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491-505.

Federmeier, K. D., Kutas, M., & Schul, R. (2010). Age-related and individual differences in the use of prediction during language comprehension. *Brain and Language, 115*(3), 149-161.

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini Meta-Analysis of Your Own Studies: Some Arguments on Why and a Primer on How. *Social and Personality Psychology Compass, 10*(10), 535-549.

Goldrick, M. A., Ferreira, V., & Miozzo, M. (2014). *The Oxford handbook of language production*. Oxford: Oxford University Press.

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience, 13*, 135-145.

Hintz, F., & Meyer, A. S. (2015). Prediction and production of simple mathematical equations: Evidence from visual world eye-tracking. *PLoS One*, *10*(7), e0130766.

Hintz, F., Meyer, A. S., & Huettig, F. (2016). Encouraging prediction during production facilitates subsequent comprehension: Evidence from interleaved object naming in sentence context and sentence reading. *Quarterly Journal of Experimental Psychology*, *69*(6), 1056-1063.

Hintz, F., Meyer, A. S., & Huettig, F. (2017). Predictors of verb-mediated anticipatory eye movements in the visual world. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(9), 1352-1374.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 118-135.

Huettig, F., & Janse, E. (2016). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. *Language, Cognition and Neuroscience, 31*(1), 80-93.

Ito, A., Corley, M., Pickering, M. J., Martin, A. E., & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language, 86*, 157-171.

Ito, A., Corley, M., & Pickering, M. J. (2017). A cognitive load delays predictive eye movements similarly during L1 and L2 comprehension. *Bilingualism: Language and Cognition*, doi:10.1017/S1366728917000050

Ito, A., Pickering, M. J., & Corley, M. (2018). Investigating the time-course of phonological prediction in native and non-native speakers of English: A visual world eye-tracking study. *Journal of Memory and Language, 98*, 1-11.

Ito, A., Gambi, C., Pickering, M. J., Fuellenbach, K., & Husband, E. M. (2020). Prediction of phonological and gender information: An event-related potential study in Italian. *Neuropsychologia, 136,* 107291.

Kukona, A., Fang, S.-Y., Aicher, K. A., Chen, H., & Magnuson, J. S. (2011). The time course of anticipatory constraint integration. *Cognition, 119*(1), 23-42.

Kukona, A. (2020). Lexical constraints on the prediction of form: Insights from the visual world paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* Advance online publication. https://doi.org/10.1037/xlm0000935

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31*(1), 32-59.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology, 62*, 621-647.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22,* 1-38.

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake-but only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance, 38*(4), 843-847.

Martin, C. D., Branzi, F. M., & Bar, M. (2018). Prediction is Production: The missing link between language production and comprehension. *Scientific Reports*, *8*(1), 1079.

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language, 66*(4), 545-567.

Nieuwland, M., Barr, D., Bartolozzi, F., Busch-Moreno, S., Donaldson, D., Ferguson, H. J., ... & Ito, A. (2019). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Proceedings of the Royal Society B: Biological Sciences*.

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., ... & Mézière, D. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, e33468.

Peterson, R. R., & Savoy, P. (1998). Lexical selection and phonological encoding during language production: Evidence for cascaded processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(3), 539-557.

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: a theory and review. *Psychological Bulletin*, *144*, 1002-1044.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences, 11*(3), 105-110.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*(04), 329-347.

Pickering, M. J., & Garrod, S. (2021). *Understanding Dialogue. Language Use and Social Interaction*. Cambridge: Cambridge University Press.

Rommers, J., Dell, G. S., & Benjamin, A. S. (2020). Word predictability blurs the lines between production and comprehension: Evidence from the production effect in memory. *Cognition, 198*, 104206.

Scott, S. K., McGettigan, C., & Eisner, F. (2009). A little more conversation, a little less action—candidate roles for the motor cortex in speech perception. *Nature Reviews Neuroscience, 10*(4), 295-302.

Stanovich, K. E., & West, R. F. (1979). Mechanisms of sentence context effects in reading: Automatic activation and conscious attention. *Memory & Cognition*, *7*(2), 77-85.

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443-467.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology, 67*, 1176-1190.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1-48.

Wicha, N. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy,

and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience*, *16*(7), 1272-1288.